Nested Unfolding Network for Real-World Concealed Object Segmentation

Chunming He[1] , Rihan Zhang[1] , Longxiang Tang[2] , Dingming Zhang[1] , Bojian Zhang[3] , Fengyang Xiao[1,†] , Jingjia Feng[1] , and Sina Farsiu[1,†]
[1]Duke University, [2]Harvard University, [3]Nankai University.

† Corresponding author. Contact: chunming.he@duke.edu

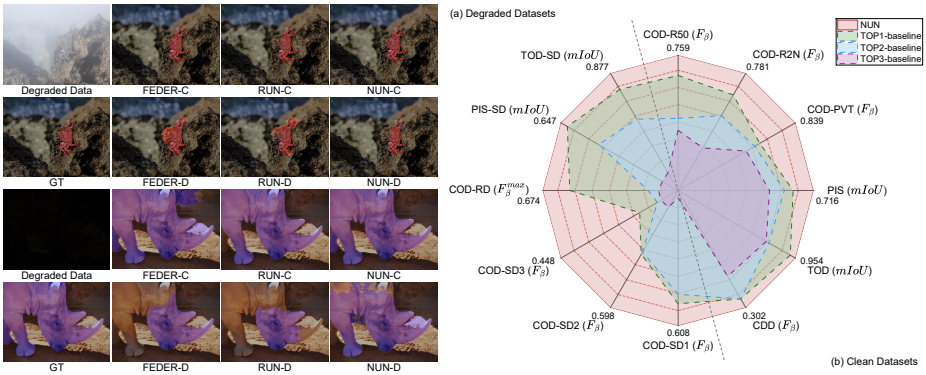

Fig. 1: Performance on COS data. **Left:** Masks are highlighted in red and purple, overlaid on the original clean data for clarity. "-C" and "-D" are shorts for clean and degraded. **Right:** The radar chart compares NUN with SOTAs over 12 COS tasks, where TOP1–TOP3 are composite baselines with the top scores per task. SD and RD denote synthetic and real-world degradation. NUN attains consistently superior results.

Abstract. Real-world visual perception demands joint handling of low-level image degradation and high-level semantic understanding, yet these two objectives inherently conflict: restoration seeks fine-grained texture recovery while segmentation prioritizes semantic contrast. We propose the nested unfolding network (NUN), a principled framework that resolves this conflict by nesting one deep unfolding network (DUN) inside another. NUN embeds a degradation-resistant unfolding network (DeRUN) within each stage of a segmentation-oriented unfolding network (SODUN), enabling both tasks to optimize in their own subspaces while interacting in a controlled manner. DeRUN handles unknown degradation through proximal gradient unfolding with learnable operators that implicitly approximate it, while SODUN performs reversible foreground-background estimation. A bi-directional unfolding interaction mechanism uses IQA to select optimal DeRUN outputs, and a cross-stage consistency loss ensures robust predictions under varying restoration quality. Theoretically, under local assumptions, we show that NUN reduces direct parameter-level gradient conflict through disjoint parameter sets and achieves a *degradation-sensitivity bound*, where degradation affects segmentation only through the inner-loop optimization and restoration

approximation errors. We instantiate NUN on concealed object segmentation and demonstrate consistent superiority over SOTA alternatives across 12 benchmarks. The code is available at https://github.com/ChunmingHe/NUN.

Keywords: Deep unfolding network · Concealed object segmentation · Real-world perception

1 Introduction

Real-world visual perception often requires jointly handling low-level image degradation and high-level semantic understanding. This challenge is especially acute in concealed object segmentation (COS) [7,14,15,17–20,62], where objects blend seamlessly into their surroundings and discriminative cues are inherently subtle. When degradations such as low light or haze further obscure these cues, the segmentation task becomes fundamentally harder (see Fig. 1).

A natural response is to combine image restoration with segmentation, yet this integration poses a core optimization conflict: restoration prioritizes fine-grained texture fidelity, while segmentation emphasizes high-level semantic contrast. Two-stage pipelines [16, 63] that first restore and then segment achieve partial degradation isolation but forfeit inter-task feedback, preventing restoration from leveraging segmentation cues. Coupled unfolding methods [21, 24] interleave both tasks within shared optimization stages, yet shared parameters cause gradient conflict between restoration and segmentation objectives, leading to compromised updates for both tasks.

We argue that what is needed is a principled architecture that enables restoration and segmentation to optimize in their own subspaces while facilitating controlled mutual refinement. Hence, we propose the nested unfolding network (NUN), which adopts a DUN-in-DUN structure: a degradation-resistant unfolding network (DeRUN) is embedded within each stage of a segmentation-oriented unfolding network (SODUN). This nested design is motivated by two insights: (1) deep unfolding provides interpretable, iterative optimization with inherent multi-stage stability, making it suitable for both tasks; and (2) the nested structure naturally creates separate optimization loops for each task while the shared multi-stage progression enables structured information exchange.

In each stage, SODUN performs reversible foreground-background estimation in both mask and RGB domains, progressively refining semantic and structural details while suppressing false positives and negatives. DeRUN, nested within each SODUN stage, performs blind degradation modeling through proximal gradient unfolding with learnable operators that implicitly approximate the unknown degradation operator, requiring no predefined degradation matrices and thereby ensuring robustness to arbitrary real-world degradation.

Leveraging the multi-stage property, we introduce the bi-directional unfolding interaction (BUI) that facilitates dynamic information exchange: segmentation outputs guide DeRUN restoration, while high-quality restorations, identified via image quality assessment, are fed back to refine subsequent segmentation

stages. To ensure coherent predictions, a cross-stage consistency constraint aligns multi-stage outputs, improving stability and robustness. We further provide theoretical analysis (Sec. 4.4) framing NUN as a principled bi-level optimization abstraction: we show that (1) coupled methods suffer from gradient conflict that nested optimization reduces at the parameter level via disjoint parameter sets, (2) the inner loop (DeRUN) converges linearly under local strong convexity, and (3) NUN admits a *degradation-sensitivity bound* where the degradation operator affects segmentation only through the inner-loop optimization error and the restoration approximation error.

Importantly, the NUN paradigm is not limited to COS: as we demonstrate through integration experiments, nesting a restoration DUN within existing segmentation frameworks consistently improves degradation robustness, suggesting promising applicability across degradation-sensitive vision tasks.

Our main contributions are summarized as follows:

(1) We introduce NUN, the first DUN-in-DUN architecture for computer vision, establishing nested unfolding as a principled paradigm for jointly optimizing low-level restoration and high-level segmentation. We provide a theoretical analysis, including an inner-loop convergence result under local assumptions and a degradation-sensitivity bound.

(2) We design SODUN for accurate segmentation in a reversible perspective and DeRUN for blind degradation modeling via proximal gradient unfolding with learnable operators, which inherently handles unknown degradation.

(3) We propose a BUI mechanism that enables controlled mutual guidance between SODUN and DeRUN through dynamic feature exchange, IQA-based selection, and cross-stage consistency, enhancing interaction and robustness.

(4) Extensive experiments across 12 benchmarks demonstrate that NUN consistently outperforms various SOTA alternatives. Integration with three different segmentation frameworks and analysis of foundation model collaboration validate its generalizability and scalability as a promising paradigm.

2 Related Works

Concealed object segmentation. Early COS methods such as SINet [9], ZoomNet [49], and FEDER [17] assume clean imaging conditions and degrade severely under real-world interference. RUN [24] introduces deep unfolding to integrate restoration with segmentation, but its coupling causes optimization conflicts and sensitivity to degradation types. RobustSAM [4] and GenSAM [29] leverage SAM [33] for improved robustness, but still struggle under severe degradation. Our NUN addresses this gap through a nested architecture that provides principled, degradation-agnostic restoration within the segmentation loop.

Deep unfolding network. DUNs unfold optimization algorithms into learnable architectures [21, 23, 24, 39, 40], achieving notable success in low-levelvision [12, 23]. In COS, RUN [24] pioneers this paradigm but assumes fixed degradation conditions. Our NUN extends DUNs through a nested architecture: an

inner module (DeRUN) performs degradation-agnostic restoration via proximal gradient unfolding, while an outer module (SODUN) carries out segmentation.
Joint restoration and perception. Combining restoration with downstream perception has been explored through two-stage pipelines [16, 59, 60], bi-level optimization [38,61,63], and coupled unfolding [21,24]. Two-stage methods treat restoration as preprocessing, yielding suboptimal outputs for downstream tasks. Bi-level optimization jointly trains both tasks but lacks structured interaction. Coupled unfolding methods interleave both tasks for iterative refinement, but shared parameters cause gradient conflict under severe degradation. Our nested unfolding provides a principled alternative, yielding consistent improvements.

3 Models and Optimization

3.1 Image segmentation

Segmentation Model. A clean image $\mathbf{X}$ can be decomposed into foreground $\mathbf{F}$ and background $\mathbf{B}$ by optimizing the following function:

$$L(\mathbf{F},\mathbf{B}) = \frac{1}{2}\|\mathbf{X}-\mathbf{F}-\mathbf{B}\|_2^2 + \beta\varphi(\mathbf{F}) + \lambda\phi(\mathbf{B}), \tag{1}$$

where $\|\cdot\|_2$ is an ℓ_2-norm. $\varphi(\cdot)$ and $\phi(\cdot)$ are regularization terms with hyper-parameters β and λ. For segmentation, the foreground is $\mathbf{F} = \mathbf{X} \odot \mathbf{M}$, with $\mathbf{M}$ and $\odot$ denoting the binary mask and dot product. Substituting into Eq. (1) gives

$$L(\mathbf{M},\mathbf{B}) = \frac{1}{2}\|\mathbf{X}-\mathbf{X}\odot\mathbf{M}-\mathbf{B}\|_2^2 + \beta\varphi(\mathbf{M}) + \lambda\phi(\mathbf{B}). \tag{2}$$

Model Optimization. Following the proximal-gradient (PG) algorithm [24], we alternately optimize $\mathbf{M}$ and $\mathbf{B}$ at stage k:

$$\mathbf{M}_k = \arg\min_{\mathbf{M}} \frac{1}{2}\|\mathbf{X}-\mathbf{X}\odot\mathbf{M}-\mathbf{B}_{k-1}\|_2^2 + \mu\psi(\mathbf{M}), \tag{3}$$

$$\mathbf{B}_k = \arg\min_{\mathbf{B}} \frac{1}{2}\|\mathbf{X}-\mathbf{X}\odot\mathbf{M}_k-\mathbf{B}\|_2^2 + \lambda\phi(\mathbf{B}). \tag{4}$$

We first optimize $\mathbf{M}_k$, containing a gradient descent term and a proximal term:

$$\hat{\mathbf{M}}_k = \mathbf{M}_{k-1} - \alpha_{\mathbf{M}}\nabla m(\mathbf{M}_{k-1}) = \mathbf{M}_{k-1} - \alpha_{\mathbf{M}}(\mathbf{X}\odot(\mathbf{X}\odot\mathbf{M}_{k-1}+\mathbf{B}_{k-1}-\mathbf{X})), \tag{5}$$

$$\mathbf{M}_k = \mathrm{prox}_{\mu,\psi}(\hat{\mathbf{M}}_k), \tag{6}$$

where $m(\cdot)$ denotes the data fidelity term of Eq. (3). $\hat{\mathbf{M}}_k$ is the intermediate solution and $\alpha_{\mathbf{M}}$ is a step-size parameter. Similarly, $\mathbf{B}_k$ is optimized with a step-size parameter $\alpha_{\mathbf{B}}$ by denoting the data fidelity term of Eq. (4) as $b(\cdot)$:

$$\hat{\mathbf{B}}_k = \mathbf{B}_{k-1} - \alpha_{\mathbf{B}}\nabla b(\mathbf{B}_{k-1}) = \mathbf{B}_{k-1} - \alpha_{\mathbf{B}}(\mathbf{B}_{k-1} + \mathbf{X}\odot\mathbf{M}_k - \mathbf{X}), \tag{7}$$

$$\mathbf{B}_k = \mathrm{prox}_{\lambda,\phi}(\hat{\mathbf{B}}_k). \tag{8}$$

3.2 Image restoration

Restoration Model. Given the degraded data $\mathbf{Y}$, the degradation process is

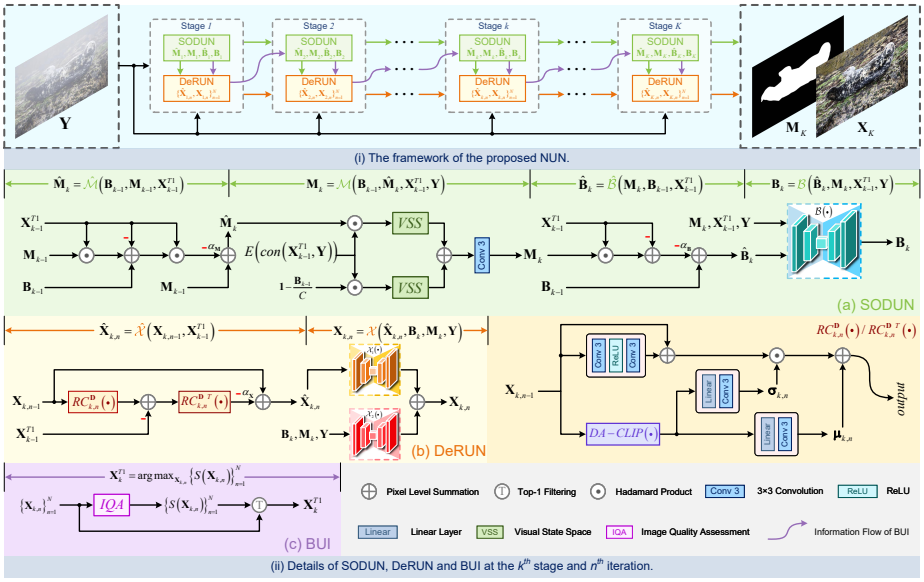

Fig. 2: Framework of our proposed NUN.

$$\mathbf{Y} = \mathbf{D}\mathbf{X} + \mathbf{N}, \tag{9}$$

where $\mathbf{D}$ denotes the unknown degradation operator and $\mathbf{N}$ is the random additive noise. The restoration objective is formulated as:

$$L(\mathbf{X}) = \frac{1}{2}\|\mathbf{Y} - \mathbf{D}\mathbf{X}\|_2^2 + \gamma\omega(\mathbf{X}), \tag{10}$$

where $\omega(\cdot)$ regularizes the restored image with weight γ.

Model Optimization. The optimization function of the restoration problem is

$$\mathbf{X}_k = \arg\min_{\mathbf{X}} \frac{1}{2}\|\mathbf{Y} - \mathbf{D}\mathbf{X}\|_2^2 + \gamma\omega(\mathbf{X}). \tag{11}$$

We use the PG algorithm to optimize the restoration problem:

$$\hat{\mathbf{X}}_k = \mathbf{X}_{k-1} - \alpha_{\mathbf{X}}\nabla x(\mathbf{X}_{k-1}) = \mathbf{X}_{k-1} - \alpha_{\mathbf{X}}\mathbf{D}^T(\mathbf{D}\mathbf{X}_{k-1} - \mathbf{Y}), \tag{12}$$

$$\mathbf{X}_k = \text{prox}_{\gamma,\omega}(\hat{\mathbf{X}}_k), \tag{13}$$

where $x(\cdot)$ is the image fidelity term of Eq. (11) and $\alpha_{\mathbf{X}}$ is the step-size parameter.

4 NUN

Motivations. Jointly optimizing restoration and segmentation within a single framework faces a fundamental challenge: the two tasks have conflicting objectives. Restoration emphasizes fine-grained texture fidelity, while segmentation prioritizes high-level semantic contrast. Coupling them in a single optimization loop, as in prior DUN-based COS methods [24], leads to gradient interference and feature entanglement. Moreover, reliance on predefined degradation types limits applicability to real-world scenes where degradation is unknown and diverse.

Algorithm 1 Proposed NUN Algorithm.

Input: degraded concealed image $\mathbf{Y}$, stage number K, iteration number N
Output: concealed object mask $\mathbf{M}_K$, restored image $\mathbf{X}_K^{T1}$

Zero initialization for $\mathbf{M}_0$, $\mathbf{B}_0$. $\mathbf{X}_0^{T1} = \mathbf{Y}$, $\mathbf{X}_{0,0} = \mathbf{Y}$
for each stage $k \in [1, K]$ **do**
  $\hat{\mathbf{M}}_k = \hat{\mathcal{M}}(\mathbf{B}_{k-1}, \mathbf{M}_{k-1}, \mathbf{X}_{k-1}^{T1})$,
  $\mathbf{M}_k = \mathcal{M}(\mathbf{B}_{k-1}, \hat{\mathbf{M}}_k, \mathbf{X}_{k-1}^{T1}, \mathbf{Y})$,
  $\hat{\mathbf{B}}_k = \hat{\mathcal{B}}\left(\mathbf{M}_k, \mathbf{B}_{k-1}, \mathbf{X}_{k-1}^{T1}\right)$,
  $\mathbf{B}_k = \mathcal{B}\left(\hat{\mathbf{B}}_k, \mathbf{M}_k, \mathbf{X}_{k-1}^{T1}, \mathbf{Y}\right)$,
  for each iteration $n \in [1, N]$ **do**
    $\hat{\mathbf{X}}_{k,n} = \hat{\mathcal{X}}(\mathbf{X}_{k,n-1}, \mathbf{X}_{k-1}^{T1})$,
    $\mathbf{X}_{k,n} = \mathcal{X}(\hat{\mathbf{X}}_{k,n}, \mathbf{B}_k, \mathbf{M}_k, \mathbf{Y})$,
  end for
  $\mathbf{X}_k^{T1} = \arg\max_{\mathbf{X}_{k,n}} \{S(\mathbf{X}_{k,n})\}_{n=1}^{N}$.
end for

We address this through a hierarchical interaction design that separates the two tasks into nested optimization loops while enabling controlled information exchange. We propose the Nested Unfolding Network (NUN), as shown in Fig. 2 and Algorithm 1. NUN is a unified, degradation-robust framework that adopts a DUN-in-DUN architecture: a degradation-resistant network (DeRUN) nested within a segmentation-oriented network (SODUN). In each stage, SODUN performs reversible foreground-background estimation, while DeRUN performs blind degradation modeling and restoration through proximal gradient unfolding with learnable operators. A bi-directional unfolding interaction (BUI) further facilitates reciprocal information exchange between segmentation and restoration. We show in Sec. 4.4 that this nested structure admits a principled bi-level optimization interpretation, with an inner-loop convergence analysis under the local assumptions stated therein.

4.1 SODUN

Derived from the segmentation model (Sec. 3.1), SODUN takes the degraded $\mathbf{Y}$ as input. Two modules, $\{\hat{\mathcal{M}}(\cdot), \mathcal{M}(\cdot)\}$ and $\{\hat{\mathcal{B}}(\cdot), \mathcal{B}(\cdot)\}$, simulate the gradient descent and proximal steps for mask prediction and background estimation.

Mask prediction. At the k^{th} stage, the intermediate mask $\mathbf{M}_k$ is obtained

$$\begin{aligned}\hat{\mathbf{M}}_k &= \hat{\mathcal{M}}(\mathbf{B}_{k-1}, \mathbf{M}_{k-1}, \mathbf{Y}),\\ &= \mathbf{M}_{k-1} - \alpha_{\mathbf{M}}(\mathbf{Y} \odot (\mathbf{Y} \odot \mathbf{M}_{k-1} + \mathbf{B}_{k-1} - \mathbf{Y})).\end{aligned} \tag{14}$$

Eq. (14) strictly follows the proximal gradient derivation, while all originally fixed parameters are made learnable to enhance generalization. With the aid of DeRUN, the enhanced result $\mathbf{X}_{k-1}$ replaces the low-quality input $\mathbf{Y}$, yielding:

$$\hat{\mathbf{M}}_k = \hat{\mathcal{M}}(\mathbf{B}_{k-1}, \mathbf{M}_{k-1}, \mathbf{X}_{k-1}). \tag{15}$$

To further refine $\hat{\mathbf{M}}$, the proximal term adopts a visual state space (VSS) module to capture non-local dependencies, which is essential for global reasoning based on sparse discriminative cues visible in concealed settings. Following common practice [24–26], we employ deep encoder features $E(\mathbf{X}_{k-1})$ extracted by

ResNet50 [28]. To mitigate potential restoration bias, both the restored and original images are utilized. The refined mask $\mathbf{M}_k$ is calculated as follows:

$$\begin{aligned}\mathbf{M}_k &= \mathcal{M}(\mathbf{B}_{k-1}, \hat{\mathbf{M}}_k, \mathbf{X}_{k-1}, \mathbf{Y}),\\ &= conv3(VSS(\hat{\mathbf{M}}_k \odot E(con(\mathbf{X}_{k-1}, \mathbf{Y})))\\ &+ VSS((\mathbf{1} - \mathbf{B}_{k-1}/\mathbf{C}) \odot E(con(\mathbf{X}_{k-1}, \mathbf{Y})))),\end{aligned} \tag{16}$$

where $con(\cdot)$, $VSS(\cdot)$, and $conv3(\cdot)$ denote concatenation, visual state space, and 3×3 convolution, respectively.

Background estimation. Given $\mathbf{M}_k$ and $\mathbf{B}_{k-1}$, the background $\hat{\mathbf{B}}_k$ is corrected via the gradient descent step, formulated as:

$$\begin{aligned}\hat{\mathbf{B}}_k &= \hat{\mathcal{B}}(\mathbf{M}_k, \mathbf{B}_{k-1}, \mathbf{X}_{k-1}),\\ &= \mathbf{B}_{k-1} - \alpha_{\mathbf{B}}(\mathbf{B}_{k-1} + \mathbf{X}_{k-1} \odot \mathbf{M}_k - \mathbf{X}_{k-1}).\end{aligned} \tag{17}$$

The proximal term refines $\hat{\mathbf{B}}_k$ using a three-layer U-shaped network $\mathcal{B}(\cdot)$ [63]:

$$\mathbf{B}_k = \mathcal{B}(\hat{\mathbf{B}}_k, \mathbf{M}_k, \mathbf{X}_{k-1}, \mathbf{Y}). \tag{18}$$

The degraded-restored pair $\{\mathbf{X}_{k-1}, \mathbf{Y}\}$ provides degradation cues that assist the proximal terms in more precise mask prediction and background correction.

Discussion. By reversibly estimating the foreground and background in both mask and RGB domains, SODUN encourages complementary feedback. This drives the model to focus on ambiguous boundary regions, mitigating false predictions and promoting segmentation results.

4.2 DeRUN

At stage k and iteration n ($1 \le n \le N$), DeRUN performs blind degradation modeling and restoration through proximal gradient unfolding. Since the degradation operator $\mathbf{D}$ in Eq. (11) is unknown in real-world scenarios, DeRUN introduces learnable residual convolutional operators $RC^{\mathbf{D}}_{k,n}(\cdot)$ and $RC^{\mathbf{D}^T}_{k,n}(\cdot)$ (with a shared structure) to implicitly approximate $\mathbf{D}$ and its transpose $\mathbf{D}^T$ [47] at each stage and iteration, formulated as:

$$RC^{\mathbf{D}}_{k,n}(\mathbf{X}) = \boldsymbol{\sigma}_{k,n} \odot (CRC(\mathbf{X}) + \mathbf{X}) + \boldsymbol{\mu}_{k,n}, \tag{19}$$

where $CRC(\cdot)$ denotes a Conv-ReLU-Conv block, and $\{\boldsymbol{\sigma}_{k,n}, \boldsymbol{\mu}_{k,n}\}$ are variational parameters that modulate the operator's behavior. In the general case, these parameters are directly learned by the network, enabling DeRUN to perform blind degradation modeling without any external prior. To further improve the approximation when semantic degradation information is available, we optionally condition $\{\boldsymbol{\sigma}_{k,n}, \boldsymbol{\mu}_{k,n}\}$ on an external degradation descriptor $\mathbf{d}_{k,n}$:

$$\boldsymbol{\sigma}_{k,n} = conv3(li(\mathbf{d}_{k,n})), \quad \boldsymbol{\mu}_{k,n} = conv3(li(\mathbf{d}_{k,n})), \tag{20}$$

with $li(\cdot)$ as a linear layer. We adopt DA-CLIP [41] as the default source of $\mathbf{d}_{k,n}$:

$$\mathbf{d}_{k,n} = DA\text{-}CLIP(\mathbf{X}_{k,n-1}), \tag{21}$$

which provides degradation semantics as a frozen, off-the-shelf module that introduces no extra parameters. Alternatively, $\{\boldsymbol{\sigma}_{k,n}, \boldsymbol{\mu}_{k,n}\}$ can be directly learned in an implicit manner (in Tab. 14), at the cost of extra learnable parameters to compensate for the absent semantic guidance. Both configurations outperform

all competing methods (Tab. 5), confirming that the core contribution lies in the nested proximal gradient structure; DA-CLIP simply offers a more parameter-efficient instantiation. The gradient update at each step is given

$$\begin{aligned}\hat{\mathbf{X}}_{k,n} &= \hat{\mathcal{X}}(\mathbf{X}_{k,n-1}, \mathbf{X}_{k-1}),\\ &= \mathbf{X}_{k,n-1} - \alpha_{\mathbf{X}} {RC^{\mathbf{D}}_{k,n}}^{T}(RC^{\mathbf{D}}_{k,n}(\mathbf{X}_{k,n-1}) - \mathbf{X}_{k-1}),\end{aligned} \tag{22}$$

where $\mathbf{X}_{k-1} = \mathbf{X}_{k-1,N}$ for brevity. This operation can be interpreted as a progressive removal of residual degradation. The proximal update is defined as:

$$\begin{aligned}\mathbf{X}_{k,n} &= \mathcal{X}(\hat{\mathbf{X}}_{k,n}, \mathbf{B}_k, \mathbf{M}_k, \mathbf{Y}),\\ &= \mathcal{X}_1(\hat{\mathbf{X}}_{k,n}) + \mathcal{X}_2(\mathbf{B}_k, \mathbf{M}_k, \mathbf{Y}),\end{aligned} \tag{23}$$

where $\mathcal{X}_1$ and $\mathcal{X}_2$ share the same structure as the lightweight network used in background estimation. We introduce the complementary term $\mathcal{X}_2$, which leverages segmentation cues to highlight critical structural information and preserve discriminative details during enhancement. This design is motivated by the observation that regions where segmentation is uncertain typically correspond to ambiguous pixel-level content; by feeding segmentation confidence information into the restoration network, DeRUN is encouraged to prioritize these challenging areas, creating a natural synergy between the two tasks.

4.3 BUI

BUI exploits the nested structure by encouraging mutual guidance. SODUN masks provide structural priors to DeRUN (via $\mathcal{X}_2$ in Eq. (23)), while DeRUN's outputs are dynamically selected to guide subsequent segmentation.

Given the intermediate outputs $\{\mathbf{X}_{k-1,n}\}_{n=1}^{N}$, we employ a comprehensive Image Quality Assessment (IQA) scheme integrating TOPIQ [2], Q-Align [58], and MUSIQ [32] to evaluate perceptual quality:

$$S(\mathbf{X}_{k-1,n}) = IQA(\mathbf{X}_{k-1,n}). \tag{24}$$

The highest quality image, $\mathbf{X}^{T1}_{k-1}$, replaces $\mathbf{X}_{k-1}$ in all SODUN and DeRUN updates (See Alg. 1). Although higher-quality inputs generally enhance segmentation, a robust framework should remain stable under minor perturbations. Hence, we introduce a cross-stage consistency loss, L_{csc}, to enforce consistent predictions between the best and second-best inputs, *i.e.*, $\mathbf{M}_k$ (superscript $T1$ omitted) and $\mathbf{M}^{T2}_k$ at stage k. The formal definition is provided in Sec. 4.5.

Robustness to error propagation. A key concern with BUI is whether early-stage mask errors mislead restoration. NUN mitigates this via: (1) IQA-based selection (Eq. (23)) that filters low-quality restorations; and (2) a CSC loss that enforces consistency between top-ranked restorations, eliminating artifacts.

4.4 Theoretical Analysis

Full proofs and detailed assumptions are in the Supp.

Bi-level formulation. NUN corresponds to a bi-level optimization: the outer level optimizes segmentation on the restored image, while the inner level optimizes restoration conditioned on segmentation variables:

$$\begin{aligned} \textit{Outer:} \quad & \min_{\mathbf{M},\mathbf{B}} \ \mathcal{L}_{\text{seg}}(\mathbf{M},\mathbf{B};\mathbf{X}^*(\mathbf{M},\mathbf{B})), \\ \textit{Inner:} \quad & \mathbf{X}^*(\mathbf{M},\mathbf{B}) = \arg\min_{\mathbf{X}} \ \mathcal{L}_{\text{res}}(\mathbf{X};\mathbf{Y},\mathbf{M},\mathbf{B}), \end{aligned} \tag{25}$$

where SODUN unrolls the outer-level and DeRUN unrolls the inner-level proximal gradient. This formulation establishes NUN as a principled framework.

Gradient conflict in coupled optimization. Coupled methods optimizing $\mathcal{L} = \mathcal{L}_{\text{seg}} + \mathcal{L}_{\text{res}}$ over shared parameters $\boldsymbol{\theta}$ suffer from gradient interference when $\cos(\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\text{seg}}, \nabla_{\boldsymbol{\theta}}\mathcal{L}_{\text{res}}) < 0$, as the cross-term $2\langle \nabla\mathcal{L}_{\text{seg}}, \nabla\mathcal{L}_{\text{res}}\rangle$ reduces the effective gradient norm. NUN reduces this conflict at the parameter level: DeRUN parameters $\boldsymbol{\theta}_{\text{res}}$ and SODUN parameters $\boldsymbol{\theta}_{\text{seg}}$ are disjoint, so each task receives an uncompromised parameter-level gradient signal. Note that indirect coupling still exists through the functional outputs $\mathbf{X}_k^{T1}$ and $\mathbf{M}_k$, which is precisely the controlled interaction that the nested structure is designed to manage.

Theorem 1 (Inner-loop convergence)**.** *Assume $\mathcal{L}_{res}$ is L_r-smooth and μ-strongly convex in $\mathbf{X}$ within a local neighborhood of $\mathbf{X}_k^*$ (see Supp. for discussion). With step size $\alpha_{\mathbf{X}} \le 1/L_r$, DeRUN converges linearly:*

$$\|\mathbf{X}_{k,N} - \mathbf{X}_k^*\| \le \left(1 - \tfrac{\mu}{L_r}\right)^{N/2} \|\mathbf{X}_{k,0} - \mathbf{X}_k^*\|, \tag{26}$$

where $\mathbf{X}_k^*$ is the inner-level minimizer at stage k. Local strong convexity is stated as an explicit assumption rather than derived; it is standard in the convergence analysis of proximal gradient methods, and over-parameterized networks empirically exhibit such behavior near convergence basins (see Supp.). The IQA-based selection chooses $\mathbf{X}_k^{T1} = \arg\max_n S(\mathbf{X}_{k,n})$ over all inner iterations; this provides an empirical tightening of the bound rather than a strictly provable improvement, since it assumes IQA scores are related to the distance to $\mathbf{X}_k^*$.

Theorem 2 (Degradation-sensitivity bound)**.** *Assume $\mathcal{L}_{seg}$ is L_0-Lipschitz continuous in $\mathbf{X}$ (i.e., $|\mathcal{L}_{seg}(\cdot;\mathbf{X}_1) - \mathcal{L}_{seg}(\cdot;\mathbf{X}_2)| \le L_0\|\mathbf{X}_1 - \mathbf{X}_2\|$). Let $\delta_k = \|\mathbf{X}_k^{T1} - \mathbf{X}_k^*\|$ denote the optimization error and $\delta_{res} = \|\mathbf{X}_k^* - \mathbf{X}\|$ denote the irreducible restoration error, where $\mathbf{X}$ is the clean image. Then:*

$$|\mathcal{L}_{\text{seg}}(\mathbf{M}_k,\mathbf{B}_k;\mathbf{X}_k^{T1}) - \mathcal{L}_{\text{seg}}(\mathbf{M}_k,\mathbf{B}_k;\mathbf{X})| \le L_0(\delta_k + \delta_{\text{res}}). \tag{27}$$

By Theorem 1, δ_k decreases exponentially with N. The term δ_{res} reflects the irreducible gap between the inner-level minimizer and the true clean image, which decreases as restoration model capacity improves.

This *degradation-sensitivity bound* means the degradation operator $\mathbf{D}$ affects segmentation only through the decomposed error $\delta_k + \delta_{\text{res}}$: δ_k is actively reduced by inner iterations and IQA selection, while δ_{res} benefits from segmentation-guided restoration. Coupled methods lack such a decomposition, directly exposing segmentation to degraded features; two-stage methods bound the segmentation gap but cannot reduce δ_k via segmentation feedback. We emphasize that this analysis is established under a surrogate formulation that treats each selected restoration $\mathbf{X}_k^{T1}$ as fixed at the corresponding stage; a fully rigorous rate for the exact bi-level problem would require additional smoothness assumptions on $\mathbf{X}^*(\mathbf{M},\mathbf{B})$, which we leave to future work.

4.5 Loss function

The total loss function is formulated as:

$$L_t = L_{basic} + \epsilon L_{csc}, \tag{28}$$

where L_{basic} is the basic loss that jointly supervises segmentation accuracy and image restoration quality. The specific definitions of L_{basic} and L_{csc} are:

$$\begin{aligned} L_{basic} = \sum_{k=1}^{K} \frac{1}{2^{K-k}} [L_{BCE}^{w}(\mathbf{M}_k, GT_s) + L_{IoU}^{w}(\mathbf{M}_k, GT_s) \\ + \|\mathbf{X}_k - \mathbf{X}\|_2^2], \end{aligned} \tag{29}$$

$$L_{csc} = \sum_{k=1}^{K} \frac{1}{2^{K-k}} [L_{BCE}^{w}(\mathbf{M}_k, \mathbf{M}_k^{T2}) + L_{IoU}^{w}(\mathbf{M}_k, \mathbf{M}_k^{T2})], \tag{30}$$

where L_{BCE}^{w} is the weighted binary cross-entropy loss and L_{IoU}^{w} is the weighted intersection-over-union loss. GT_s and $\mathbf{X}$ denote the ground truth of the mask and clean data. We use $\mathbf{X}_k$ instead of $\mathbf{X}_k^{T1}$ to implicitly encourage the network to generate higher-quality outputs as the iterations progress.

Our loss enables NUN to learn degradation-invariant representations while preserving fine-grained structural fidelity and ensuring stable training.

5 Experiment

Experimental setup. Our NUN is implemented in PyTorch and trained on two NVIDIA RTX 4090 GPUs using Adam with momentum terms (0.9, 0.999). Following the common practice [9, 24], we incorporate multi-level deep features extracted from encoder-shaped backbones into our framework. All input images are resized to 352×352. The batch size is set to 36, and the initial learning rate is 1×10^{-4}, decreased by a factor of 0.1 every 80 epochs. The stage number K of SODUN is fixed at 4. For the nested DeRUN, its stage number N varies with K, *i.e.*, $N \in \{4, 3, 3, 2\}$ across the four stages.

5.1 Clean Datasets and Metrics

Camouflaged object detection. Following SINet [9], we use four benchmarks: *CHAMELEON* [51] (76 images), *CAMO* [34] (1,250 images), *COD10K* [7] (5,066 images, ten super-classes), and *NC4K* [43] (4,121 images, the largest test set). Training uses 1,000 *CAMO* and 3,040 *COD10K* images; the remaining samples, together with all *CHAMELEON* and *NC4K* images, are used for testing. We adopt four metrics: mean absolute error (M), adaptive F-measure (F_β) [44], mean E-measure (E_ϕ) [8], and structure measure (S_α) [6], where lower M and higher F_β, E_ϕ, S_α indicate better results.

Polyp image segmentation. For polyp image segmentation (PIS), we experiment on *CVC-ColonDB* [54] and *ETIS* [50] following LSSNet [55], reporting mean Dice (mDice), mean IoU (mIoU), and structure measure (S_α), where higher is better. To ensure fair comparison with SOTA medical methods that typically use transformer-based encoders, we adopt PVT-V2 as the backbone.

Transparent object detection. For transparent object detection (TOD), we also adopt PVT-V2 and evaluate on *GDD* [45] and *GSD* [37], training on 2,980 *GDD* and 3,202 *GSD* images with the rest for testing. Following GDNet-B [46], we report mIoU and maximum F-measure (F_β^{max}), where higher values reflect superior performance.
Concealed defect detection. For concealed defect detection (CDD), we assess zero-shot generalization by directly applying the COD-trained model, without fine-tuning, to *CDS2K* [10], with PVT-V2 as the backbone. We employ eight metrics, where higher values are better except for MAE.

5.2 Degraded COS Datasets

We build a comprehensive real-world benchmark from two origins: synthesized corruptions on clean datasets, and empirical low-quality subsets filtered from existing collections.
Simulated corruptions. We synthesize three corruption types critical for concealed object tasks: low-light and haze, which actively obscure subtle discriminative cues, and low-resolution, which passively reduces information fidelity.

- Low-light: the Retinex-based strategy of Zhang et al. [64], simulating illumination attenuation and noise amplification to yield low-SNR images.
- Haze: the atmospheric scattering model following Ancuti et al. [35], spanning light to heavy fog.
- Low-resolution: the practical degradation model of Zhang et al. [65], integrating Gaussian-kernel blur, downsampling, and noise injection.

All corruptions span multiple severity levels (light, medium, heavy). To match application scenarios, polyp images receive low-resolution degradation, indoor transparent objects receive low-light, and outdoor transparent objects receive combined degradations.
Empirical low-quality subsets. To assess real-world generalization to unseen corruptions, we build two low-quality subsets, *PCOD-LQ* and *MCOD-LQ*, filtered from *PCOD* [57] and *MCOD* [36]. Four annotators independently flagged visually degraded samples (e.g., uneven illumination, blurring) without referencing degradation types; only images with $\geq$3-vote consensus were retained, yielding 160 samples (*PCOD-LQ*: 90; *MCOD-LQ*: 70). They naturally contain out-of-distribution corruptions such as overexposure and snowfall.

5.3 COS tasks with clean data

Datasets and metrics are in the Supp. For clean data, DeRUN only acts as an image reconstruction module using estimated foreground and background.
Camouflaged object detection. Following FEDER [17], we assess the performance of our model on COD under four backbones: ResNet50 [28], Res2Net50 [13], PVT V2 [56], and DINO v2 [48]. To ensure fair comparison, within each

Table 1: Results on camouflaged object detection with clean data.

Methods	Backbones	*CHAMELEON*				*CAMO*				*COD10K*				*NC4K*			
		$M\downarrow$	$F_\beta\uparrow$	$E_\phi\uparrow$	$S_\alpha\uparrow$	$M\downarrow$	$F_\beta\uparrow$	$E_\phi\uparrow$	$S_\alpha\uparrow$	$M\downarrow$	$F_\beta\uparrow$	$E_\phi\uparrow$	$S_\alpha\uparrow$	$M\downarrow$	$F_\beta\uparrow$	$E_\phi\uparrow$	$S_\alpha\uparrow$
FEDER [17]	ResNet50	0.028	0.850	0.944	0.892	0.070	0.775	0.870	0.802	0.032	0.715	0.892	0.810	0.046	0.808	0.900	0.842
FSEL [53]	ResNet50	0.029	0.847	0.941	0.893	**0.069**	0.779	0.881	**0.816**	0.032	0.722	0.891	0.822	0.045	0.807	0.901	0.847
RUN [24]	ResNet50	0.027	0.855	0.952	0.895	0.070	0.781	0.868	0.806	0.030	0.747	0.903	0.827	**0.042**	0.824	0.908	0.851
NUN (Ours)	ResNet50	**0.026**	**0.862**	**0.955**	**0.899**	**0.069**	**0.788**	**0.890**	**0.818**	**0.029**	**0.759**	**0.908**	**0.831**	**0.042**	**0.828**	**0.913**	**0.855**
BSA-Net [66]	Res2Net50	0.027	0.851	0.946	0.895	0.079	0.768	0.851	0.796	0.034	0.723	0.891	0.818	0.048	0.805	0.897	0.841
RUN [24]	Res2Net50	0.024	0.879	0.956	0.907	**0.066**	0.815	0.905	0.843	0.028	0.764	0.914	0.849	0.041	0.830	0.917	0.859
NUN (Ours)	Res2Net50	**0.023**	**0.883**	**0.959**	**0.910**	**0.066**	**0.822**	**0.909**	**0.846**	**0.026**	**0.781**	**0.920**	**0.852**	**0.040**	**0.836**	**0.922**	**0.864**
CamoDiff [52]	PVT V2	0.022	0.868	0.952	0.908	**0.042**	0.853	0.936	0.878	0.019	0.815	0.943	0.883	0.028	0.858	0.942	0.895
RUN [24]	PVT V2	**0.021**	0.877	0.958	0.916	0.045	0.861	0.934	0.877	0.021	0.810	0.941	0.878	0.030	0.868	0.940	0.892
NUN (Ours)	PVT V2	**0.021**	**0.888**	**0.963**	**0.918**	**0.042**	**0.868**	**0.939**	**0.885**	**0.018**	**0.839**	**0.948**	**0.890**	**0.027**	**0.880**	**0.947**	**0.897**
VSCode [42]	DINOv2-L	0.020	0.882	0.954	0.915	0.043	0.860	0.931	0.882	0.020	0.818	0.935	0.880	0.030	0.867	0.940	0.895
C3Net [31]	DINOv2-L	0.016	0.905	0.968	0.927	0.031	0.892	0.953	0.904	0.016	0.852	0.961	0.898	0.024	0.894	0.958	0.913
NUN (Ours)	*DINOv2-L*	**0.015**	**0.918**	**0.975**	**0.931**	**0.028**	**0.908**	**0.965**	**0.911**	**0.014**	**0.880**	**0.972**	**0.905**	**0.020**	**0.913**	**0.967**	**0.920**

Table 2: Results on clean TOD (*GDD* [45]).

Methods	mIoU ↑	$F_\beta^{max}\uparrow$	$M\downarrow$
EBLNet [27]	0.870	0.922	0.064
RFENet [11]	0.886	0.938	0.057
RUN [24]	0.895	0.952	0.051
NUN (Ours)	**0.903**	**0.956**	**0.048**

Table 3: Results on clean PIS (*ETIS* [50]).

Methods	mDice ↑	mIoU ↑	$S_\alpha\uparrow$
PolypPVT [5]	0.787	0.706	0.871
MEGANet [1]	0.739	0.665	0.836
RUN [24]	0.788	0.709	0.878
NUN (Ours)	**0.791**	**0.716**	**0.882**

Table 4: Results on clean CDD (*CDS2K* [10]).

Methods	$S_\alpha\uparrow$	$M\downarrow$	$E_\phi\uparrow$	$F_\beta\uparrow$
HitNet [30]	0.563	0.118	0.564	0.298
FEDER [17]	0.538	0.070	0.586	0.288
RUN [24]	0.590	0.068	0.595	0.298
NUN (Ours)	**0.595**	**0.066**	**0.599**	**0.303**

Table 5: Results on real-world COD with synthetic degradation data.

Methods	Backbones	*CHAMELEON*				*CAMO*				*COD10K*				*NC4K*			
		$M\downarrow$	$F_\beta\uparrow$	$E_\phi\uparrow$	$S_\alpha\uparrow$	$M\downarrow$	$F_\beta\uparrow$	$E_\phi\uparrow$	$S_\alpha\uparrow$	$M\downarrow$	$F_\beta\uparrow$	$E_\phi\uparrow$	$S_\alpha\uparrow$	$M\downarrow$	$F_\beta\uparrow$	$E_\phi\uparrow$	$S_\alpha\uparrow$
Single Degradation: Low-Light																	
FEDER [17]	ResNet50	0.056	0.721	0.847	0.817	0.130	0.537	0.693	0.636	0.066	0.510	0.742	0.670	0.092	0.619	0.764	0.698
FSEL [53]	ResNet50	0.134	0.221	0.424	0.485	0.175	0.217	0.417	0.462	0.092	0.213	0.518	0.513	0.146	0.253	0.463	0.492
RUN [24]	ResNet50	0.057	0.729	0.841	0.813	0.124	0.579	0.737	0.657	0.063	0.539	0.785	0.684	0.089	0.643	0.795	0.713
NUN (Ours)	ResNet50	**0.054**	**0.736**	**0.851**	**0.824**	**0.112**	**0.634**	**0.754**	**0.724**	**0.055**	**0.608**	**0.795**	**0.751**	**0.072**	**0.714**	**0.826**	**0.794**
Single Degradation: Hazy																	
FEDER [17]	ResNet50	0.074	0.641	0.744	0.753	0.134	0.543	0.677	0.656	0.060	0.525	0.722	0.728	0.082	0.675	0.746	0.754
FSEL [53]	ResNet50	0.075	0.643	0.755	0.751	0.144	0.505	0.668	0.623	0.081	0.467	0.721	0.656	0.102	0.596	0.752	0.696
RUN [24]	ResNet50	0.067	0.652	0.769	0.748	0.128	0.547	0.681	0.636	0.059	0.528	0.726	0.731	0.079	0.678	0.754	0.764
NUN (Ours)	ResNet50	**0.059**	**0.706**	**0.815**	**0.805**	**0.120**	**0.574**	**0.704**	**0.685**	**0.053**	**0.598**	**0.780**	**0.743**	**0.072**	**0.699**	**0.811**	**0.782**
Combined Degradation: Low-Resolution, Low-Light, and Hazy																	
FEDER [17]	ResNet50	0.129	0.422	0.669	0.563	0.169	0.391	0.591	0.529	0.104	0.322	0.657	0.539	0.147	0.408	0.645	0.554
FSEL [53]	ResNet50	0.136	0.413	0.676	0.559	0.173	0.394	0.579	0.531	0.110	0.317	0.637	0.547	0.150	0.397	0.649	0.547
RUN [24]	ResNet50	0.125	0.457	0.680	0.574	0.169	0.394	0.578	0.533	0.100	0.345	0.664	0.548	0.143	0.547	0.630	0.572
NUN (Ours)	ResNet50	**0.071**	**0.629**	**0.750**	**0.729**	**0.137**	**0.463**	**0.619**	**0.599**	**0.068**	**0.448**	**0.686**	**0.643**	**0.096**	**0.572**	**0.716**	**0.681**
RUN [24]	PVT V2	0.085	0.599	0.800	0.714	0.142	0.496	0.659	0.614	0.085	0.429	0.712	0.627	0.115	0.543	0.72	0.657
NUN (Ours)	PVT V2	**0.028**	**0.862**	**0.943**	**0.890**	**0.077**	**0.770**	**0.846**	**0.784**	**0.035**	**0.732**	**0.880**	**0.808**	**0.048**	**0.815**	**0.898**	**0.838**

backbone group all methods (including ours) use identical backbone, input resolution, and training schedule; heavier encoders (PVT V2, DINOv2-L) are compared only against baselines using the same encoder, so reported gains reflect the nested-unfolding design rather than backbone capacity. As shown in Tab. 1, our NUN shows superior robustness and generalization capability across all four datasets.

Transparent object detection. As illustrated in Tab. 2, our NUN surpasses existing methods across all datasets, highlighting its potential to advance vision-based perception systems for autonomous driving.

Polyp image segmentation. We evaluate NUN on PIS using the *ETIS* dataset. Following RUN [24], PVT V2 is adopted as the default encoder. As shown in Tab. 3, our NUN consistently achieves leading performance across all datasets.

Concealed defect detection. We assess the zero-shot generalizability to CDD, with models pre-trained on COD applied to *CDS2K* to segment defects. Adhering to [24], PVT V2 serves as the default backbone. As verified by Tab. 4, our NUN achieves superior performance than others.

Table 6: Results on real-world degraded COD.

Methods	*PCOD-LQ*			*MCOD-LQ*		
	mIoU ↑	F_β^{max} ↑	M ↓	mIoU ↑	F_β^{max} ↑	M ↓
FEDER [17]	0.104	0.585	0.481	0.159	0.440	0.438
FSEL [53]	0.108	0.593	0.463	0.230	0.338	0.453
RUN [24]	0.102	0.653	0.499	0.045	0.490	0.480
NUN (Ours)	**0.061**	**0.674**	**0.657**	**0.034**	**0.556**	**0.522**

Table 7: Results on degraded TOD.

Methods	*GDD*		
	mIoU ↑	F_β^{max} ↑	M ↓
EBLNet [27]	0.786	0.873	0.108
RFENet [11]	0.835	0.896	0.081
RUN [24]	0.862	0.920	0.070
NUN (Ours)	**0.877**	**0.928**	**0.062**

Table 8: Results on degraded PIS.

Methods	*ETIS*		
	mDice ↑	mIoU ↑	S_α ↓
PolypPVT [5]	0.580	0.511	0.731
MEGANet [1]	0.667	0.596	0.810
RUN [24]	0.715	0.638	0.830
NUN (Ours)	**0.722**	**0.647**	**0.836**

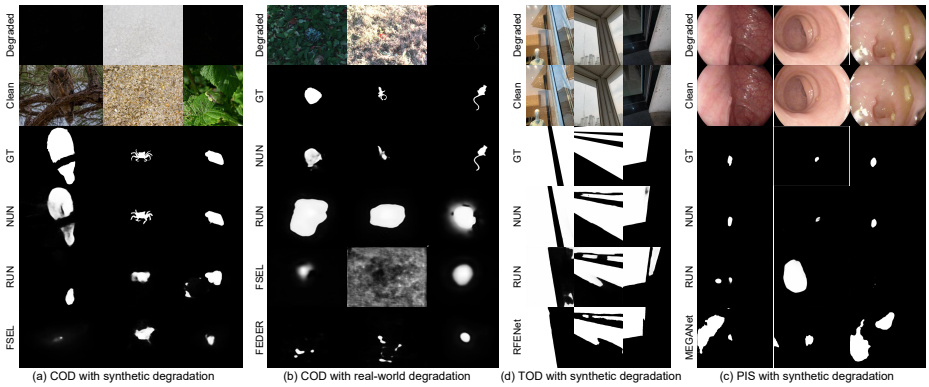

Fig. 3: Visualization on COS tasks with real-world and synthetic degradation.

Table 9: Ablations of NUN.

SODUN-	SODUN	DeRUN	BUI	M ↓	F_β ↑	E_ϕ ↑	S_α ↑
✓	×	×	×	0.102	0.328	0.603	0.583
✓	✓	×	×	0.093	0.362	0.619	0.592
✓	✓	✓	×	0.076	0.406	0.646	0.618
✓	✓	✓	✓	**0.068**	**0.448**	**0.686**	**0.643**

Table 10: Ablation study of SODUN.

Metrics	$\mathcal{M}_1(\cdot)\to\mathcal{M}(\cdot)$	$\mathcal{M}_1(\cdot)\to\mathcal{M}(\cdot)$	$\mathcal{B}_1(\cdot)\to\mathcal{B}(\cdot)$	$\mathcal{B}_2(\cdot)\to\mathcal{B}(\cdot)$	NUN (Ours)
M ↓	0.074	0.073	**0.068**	0.069	**0.068**
F_β ↑	0.412	0.416	0.447	0.445	**0.448**
E_ϕ ↑	0.663	0.666	**0.687**	0.680	0.686
S_α ↑	0.630	0.628	0.641	0.639	**0.643**

5.4 COS tasks with degraded data

COD with synthetic degradation. We simulate real-world conditions using three degradations (low-light, hazy, low-resolution) on four datasets [16, 23] (see Supp) and retrain all methods. As shown in Tab. 5 and Fig. 3, FEDER [17] and FSEL [53] degrade notably, while our NUN remains robust. NUN consistently outperforms the second-best method, RUN [24], by 7.65%, 6.90%, and 20.42% under low-light, hazy, and combined degradations. Using the stronger PVT-V2 backbone yields further gains, underscoring NUN's practicality.

COD with real degradation. To evaluate real-world generalization, we construct two low-quality subsets: *PCOD-LQ* and *MCOD-LQ*, filtered from *PCOD* [57] and *MCOD* [36]. Degraded samples are selected by four experts with ≥3-vote consensus, yielding 160 images (*PCOD-LQ*: 90; *MCOD-LQ*: 70) that include out-of-distribution degradations (e.g., over-exposure, snowfall). Models trained on COD data with combined synthetic degradations are directly evaluated. As shown in Tab. 6 and Fig. 3, NUN performs best, demonstrating strong real-world generalization and validating the realism of our synthetic degradations.

Other COS tasks with synthetic degradation. We further extend our evaluation by introducing synthetic degradations into other COS tasks. We add different types of degradation to different tasks, accommodating their specific scenarios (see Supp.). All competing methods are retrained. As reported in Tabs. 7 and 8 and Fig. 3, our NUN achieves the best performance across all datasets, demonstrating its robustness and adaptability to diverse degradation scenarios.

Table 11: Ablation study of DeRUN and BUI.

Metrics	DeRUN			BUI			NUN
	$\mathcal{X}_1(\cdot)\to\hat{\mathcal{X}}(\cdot)$	$\mathcal{X}_3(\cdot)\to(\mathcal{X}_1(\cdot)\&\mathcal{X}_2(\cdot))$	w/o $\mathcal{X}_2(\cdot)$	w/o $IQA(\cdot)$	$IQA_1(\cdot)\to IQA(\cdot)$	w/o L_{csc}	(Ours)
$M\downarrow$	0.073	0.068	0.071	0.072	0.070	0.071	**0.068**
$F_\beta\uparrow$	0.422	0.452	0.430	0.421	0.435	0.428	**0.448**
$E_\phi\uparrow$	0.659	0.685	0.671	0.655	0.670	0.665	**0.686**
$S_\alpha\uparrow$	0.628	0.640	0.633	0.626	0.636	0.633	**0.643**

Table 12: Rest. LQ.

Methods	PSNR ↑	FID ↓	User Study
Degraded	11.17	95.63	1.17
DiffIR [59]	18.63	81.07	3.58
RUN [24]	15.21	90.12	2.25
DeRUN	18.04	86.21	3.08
NUN	**22.37**	**73.15**	**3.92**

Table 13: Comparison with SAM-based methods on *COD10K*.

Methods	Backbone	Clean		Degraded	
		$F_\beta\uparrow$	$S_\alpha\uparrow$	$F_\beta\uparrow$	$S_\alpha\uparrow$
RobustSAM [4]	SAM-ViT-H	0.862	0.877	0.669	0.622
SAMadapter [3]	SAM-ViT-H	0.813	0.883	0.512	0.528
NUN (Ours)	ViT-H	**0.903**	**0.912**	**0.806**	**0.733**

Table 14: Different degradation prior configurations. UB is short for upper bound.

Configurations	$M\downarrow$	$F_\beta\uparrow$	$S_\alpha\uparrow$	Time ↓
w/o VLM (implicit learning)	0.070	0.440	0.637	**30ms**
VLM @ stage 1 only (reuse, default)	0.068	0.448	0.643	38ms
VLM @ every iteration	**0.067**	**0.457**	**0.647**	44ms
Known degradation types (UB)	0.062	0.486	0.676	22ms

5.5 Ablation Study

Unless otherwise specified, experiments of Secs. 5.5 and 5.6 are conducted on *COD10K* under combined synthetic degradations with the default backbone of ResNet50. Please see the ablations for the stage numbers in the Supp.

Effect of SODUN. As shown in Tab. 9, removing the background estimation part (SODUN-) leads to performance drop. Furthermore, in Tab. 10, we analyze the impact of different architectural modifications: (1) adding extra regularization on Eq. (2) like RUN [24] ($\hat{\mathcal{M}}_1(\cdot)\to\hat{\mathcal{M}}(\cdot)$), changing the VSS module used in Eq. (16) with a CNN module with similar parameters ($\mathcal{M}_1(\cdot)\to\mathcal{M}(\cdot)$), replacing our simple network with a complex transformer [16] in Eq. (18) ($\mathcal{B}_1(\cdot)\to\mathcal{B}(\cdot)$), and removing $\mathbf{Y}$ in Eq. (18) to break the clean-degradation pair. The performance decreases indicate the effect of our SODUN.

Effect of DeRUN. The effect of DeRUN is jointly validated in Tabs. 9 and 11. As shown in Tab. 11, several modifications lead to degraded performance: (1) replacing the explicit prior from DA-CLIP in Eq. (22) with implicit prior learned by the networks ($\hat{\mathcal{X}}_1(\cdot)\to\hat{\mathcal{X}}(\cdot)$); (2) replacing the light-weight network in Eq. (23) with more complex DiffIR used by RUN [24] ($\mathcal{X}_3(\cdot)\to(\mathcal{X}_1(\cdot)\&\mathcal{X}_2(\cdot))$); and (3) removing the segmentation cues in Eq. (23) (w/o $\mathcal{X}_2(\cdot)$). These consistent performance declines highlight the rationality and effectiveness of our DeRUN.

Effect of BUI. The BUI module is designed to enhance information exchange within the DUN-in-DUN architecture. As shown in Tabs. 9 and 11, both components, the IQA-based data selection strategy ($IQA_1(\cdot)$, which relies solely on MUSIQ for evaluation, similar to CORUN [12]) and the consistency loss, contribute to performance improvement. They verify the superiority of our BUI.

5.6 Further Analysis and Applications

In the future, we will explore the potential of such a nested framework in more settings, such as incomplete supervision [19, 22].

Performance on unseen degradation types. Trained on synthetic degradations, NUN generalizes well to real-world and unseen degradations (e.g., overexposure, snow), achieving robust performance in Tab. 6.

Image restoration quality. As shown in Tab. 12, NUN achieves top results, with a user study containing 12 subjects. In degraded COS, RUN [24] integrates DiffIR [59] as its restoration backbone, yet its performance drops due to conflicts between restoration and segmentation goals. In contrast, NUN alleviates such

conflicts and surpasses both DiffIR and the standalone DeRUN, confirming its ability to jointly enhance restoration and segmentation.

Comparison with SAM-based methods. As shown in Tab. 13, NUN outperforms SAM-based models on COD10K under both clean and degraded settings. While gains are moderate on clean data, the gap widens under degradation, showing the superiority of our framework.

Degradation prior configurations. Tab. 14 shows three DeRUN variants with increasing accuracy: implicit learning (directly learning $\{\boldsymbol{\sigma}_{k,n}, \boldsymbol{\mu}_{k,n}\}$), VLM @ stage 1 (default), and VLM @ every iteration. Even implicit learning surpasses all SOTA methods, indicating the effect of DeRUN's degradation modeling.

Toward a unified degradation-robust vision paradigm. The progression from two-stage through BLO to nested unfolding confirms that tighter restoration-perception integration yields better performance. NUN simultaneously improves both tasks (Tab. 12). The bi-level formulation (Eq. (25)) is not specific to COS, suggesting promising applicability to other degradation-sensitive tasks such as adverse-weather detection and noisy medical imaging. The modular structure further allows incorporating foundation model backbones without altering the stated bound (Eq. (27) and Tab. 1).

6 Conclusions

We propose nested unfolding as a principled framework for jointly optimizing image restoration and segmentation, instantiated as NUN for real-world COS. The resulting DUN-in-DUN architecture mitigates the inherent conflict between low- and high-level vision objectives by formulating joint learning as a BLO problem, providing an inner-loop convergence analysis under local assumptions, reducing parameter-level gradient interference, and bounding degradation effects on segmentation through a controllable restoration error. Extensive experiments on 12 benchmarks demonstrate our superiority and generalization.

Acknowledgements

This work was supported in part by the Foundation Fighting Blindness (BR-CL-0621-0812-DUKE and PPA-1224-0890-DUKE) and a Research to Prevent Blindness Unrestricted Grant to Duke University.

References

1. Bui, N.T., Hoang, D.H., Nguyen, Q.T., Tran, M.T., Le, N.: Meganet: Multi-scale edge-guided attention network for weak boundary polyp segmentation. In: WACV. pp. 7985–7994 (2024) 12, 13
2. Chen, C., Mo, J., Hou, J., Wu, H., Liao, L., Sun, W., Yan, Q., Lin, W.: Topiq: A top-down approach from semantics to distortions for image quality assessment. IEEE Transactions on Image Processing **33**, 2404–2418 (2024) 8

3. Chen, T., Zhu, L., Deng, C., Cao, R., Wang, Y., Zhang, S.: Sam-adapter: Adapting segment anything in underperformed scenes. In: ICCV. pp. 3367–3375 (2023) 14
4. Chen, W.T., Vong, Y.J., Kuo, S.Y., Ma, S., Wang, J.: Robustsam: Segment anything robustly on degraded images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4081–4091 (2024) 3, 14
5. Dong, B., Wang, W., Fan, D.P., Li, J., Fu, H., Shao, L.: Polyp-pvt: Polyp segmentation with pyramid vision transformers. CAAI Artif. Intell. Res. **2** (2023) 12, 13
6. Fan, D.P., Cheng, M.M., Liu, Y., Li, T.: Structure-measure: A new way to evaluate foreground maps. In: ICCV. pp. 4548–4557 (2017) 10
7. Fan, D.P., Ji, G.P., Cheng, M.M., Shao, L.: Concealed object detection. IEEE Trans. Pattern Anal. Mach. Intell. (2021) 2, 10
8. Fan, D.P., Ji, G.P., Qin, X., Cheng, M.M.: Cognitive vision inspired object segmentation metric and loss function. Scientia Sinica Informationis **6**(6) (2021) 10
9. Fan, D.P., Ji, G.P., Sun, G., Cheng, M.M., Shen, J.: Camouflaged object detection. In: CVPR. pp. 2777–2787 (2020) 3, 10
10. Fan, D.P., Ji, G.P., Xu, P., Cheng, M.M., Sakaridis, C., Van Gool, L.: Advances in deep concealed scene understanding. Visual Intell. **1**(1), 16 (2023) 11, 12
11. Fan, K., Wang, C., Wang, Y., Wang, C., Yi, R., Ma, L.: Rfenet: towards reciprocal feature evolution for glass segmentation. In: IJCAI. pp. 717–725 (2023) 12, 13
12. Fang, C., He, C., Xiao, F., Zhang, Y., Tang, L., Zhang, Y., Li, K., Li, X.: Real-world image dehazing with coherence-based pseudo labeling and cooperative unfolding network. In: The Thirty-eighth Annual Conference on Neural Information Processing Systems (2024) 3, 14
13. Gao, S.H., Cheng, M.M., Zhao, K., Zhang, X.Y., Yang, M.H., Torr, P.: Res2net: A new multi-scale backbone architecture. IEEE Trans. Pattern Anal. Mach. Intell. **43**(2), 652–662 (2019) 11
14. Gao, Z., Jia, W., Zhang, X., Zhou, D., Xu, K., Dawei, F., Dou, Y., Mao, X., Wang, H.: Knowledge memorization and rumination for pre-trained model-based class-incremental learning. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 20523–20533 (2025) 2
15. Gao, Z., Sun, Z., Zhang, X., Xu, K., Wang, H.: Re-evaluating continual vqa: Toward fair and robust evaluation for multimodal continual learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 18021–18031 (June 2026) 2
16. He, C., Fang, C., Zhang, Y., Ye, T., Li, K., Tang, L., Guo, Z., Li, X., Farsiu, S.: Reti-diff: Illumination degradation image restoration with retinex-based latent diffusion model. ICLR (2025) 2, 4, 13, 14
17. He, C., Li, K., Zhang, Y., Tang, L., Zhang, Y.: Camouflaged object detection with feature decomposition and edge reconstruction. In: CVPR. pp. 22046–22055 (2023) 2, 3, 11, 12, 13
18. He, C., Li, K., Zhang, Y., Xu, G., Tang, L.: Weakly-supervised concealed object segmentation with sam-based pseudo labeling and multi-scale feature grouping. NeurIPS (2024) 2
19. He, C., Li, K., Zhang, Y., Yang, Z., Tang, L., Zhang, Y., Kong, L., Farsiu, S.: Segment concealed object with incomplete supervision. IEEE Trans. Pattern Anal. Mach. Intell. (2025) 2, 14
20. He, C., Li, K., Zhang, Y., Zhang, Y., Guo, Z., Li, X.: Strategic preys make acute predators: Enhancing camouflaged object detectors by generating camouflaged objects. ICLR (2024) 2

21. He, C., Xiao, F., Zhang, R., Fang, C., Fan, D.P., Farsiu, S.: Reversible unfolding network for concealed visual perception with generative refinement. arXiv preprint arXiv:2508.15027 (2025) 2, 3, 4
22. He, C., Zhang, R., Tang, L., Yang, Z., Li, K., Fan, D.P., Farsiu, S.: Scaler: Sam-enhanced collaborative learning for label-deficient concealed object segmentation. arXiv preprint arXiv:2511.18136 (2025) 14
23. He, C., Zhang, R., Xiao, F., Fang, C., Tang, L., Zhang, Y., Farsiu, S.: Unfoldir: Rethinking deep unfolding network in illumination degradation image restoration. CVPR (2026) 3, 13
24. He, C., Zhang, R., Xiao, F., Fang, C., Tang, L., Zhang, Y., Kong, L., Fan, D.P., Li, K., Farsiu, S.: Run: Reversible unfolding network for concealed object segmentation. ICML (2025) 2, 3, 4, 5, 6, 10, 12, 13, 14
25. He, C., Zhang, R., Xiao, F., Zhang, D., Cao, Z., Farsiu, S.: Refining context-entangled content segmentation via curriculum selection and anti-curriculum promotion. In: ICML (2026) 6
26. He, C., Zhang, R., Zhang, D., Fang, C., Tang, L., Feng, J., Xiao, F., Farsiu, S.: Ride: Retinex-informed decoupling for exposing concealed objects. arXiv preprint arXiv:2605.15450 (2026) 6
27. He, H., Li, X., Cheng, G., Shi, J., Tong, Y., Meng, G., Prinet, V., Weng, L.: Enhanced boundary learning for glass-like object segmentation. In: ICCV. pp. 15859–15868 (2021) 12, 13
28. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016) 7, 11
29. Hu, J., Lin, J., Gong, S.: Relax image-specific prompt requirement in sam: A single generic prompt for segmenting camouflaged objects. In: AAAI. vol. 38, pp. 12511–12518 (2024) 3
30. Hu, X., Wang, S., Qin, X., Dai, H., Ren, W., Luo, D., Tai, Y., Shao, L.: High-resolution iterative feedback network for camouflaged object detection. In: AAAI. vol. 37, pp. 881–889 (2023) 12
31. Jan, B., El-Maleh, A.H., Siddiqui, A.J., Bais, A., Anwar, S.: C3net: Context-contrast network for camouflaged object detection. arXiv preprint arXiv:2511.12627 (2025) 12
32. Ke, J., Wang, Q., Wang, Y., Milanfar, P., Yang, F.: Musiq: Multi-scale image quality transformer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5148–5157 (2021) 8
33. Kirillov, A., Mintun, E., Ravi, N., Mao, H.: Segment anything. arXiv preprint arXiv:2304.02643 (2023) 3
34. Le, T.N., Nguyen, T.V., Nie, Z., Tran, M.T., Sugimoto, A.: Anabranch network for camouflaged object segmentation. Comput. Vis. Image Underst. **184**, 45–56 (2019) 10
35. Li, B., Ren, W., Fu, D., Tao, D., Feng, D., Zeng, W., Wang, Z.: Benchmarking single image dehazing and beyond (2019), https://arxiv.org/abs/1712.04143 11
36. Li, Y., Xu, T., Bai, S., Liu, P., Li, J.: Mcod: The first challenging benchmark for multispectral camouflaged object detection. arXiv preprint arXiv:2509.15753 (2025) 11, 13
37. Lin, J., He, Z.: Rich context aggregation with reflection prior for glass surface detection. In: CVPR. pp. 13415–13424 (2021) 11
38. Liu, R., Liu, Y., Zeng, S., Zhang, J.: Augmenting iterative trajectory for bilevel optimization: Methodology, analysis and extensions. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025) 4

39. Lu, S., Liu, Y., Kong, A.W.K.: Tf-icon: Diffusion-based training-free cross-domain image composition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2294–2305 (2023) 3
40. Lu, S., Wang, Z., Li, L., Liu, Y., Kong, A.W.K.: Mace: Mass concept erasure in diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6430–6440 (2024) 3
41. Luo, Z., Gustafsson, F.K., Zhao, Z., Sjölund, J., Schön, T.B.: Controlling vision-language models for universal image restoration. arXiv preprint arXiv:2310.01018 (2023) 7
42. Luo, Z., Liu, N., Zhao, W., Yang, X., Zhang, D., Fan, D.P., Khan, F., Han, J.: Vscode: General visual salient and camouflaged object detection with 2d prompt learning. In: CVPR. pp. 17169–17180 (2024) 12
43. Lv, Y., Zhang, J., Dai, Y., Li, A., Liu, B., Barnes, N., Fan, D.P.: Simultaneously localize, segment and rank the camouflaged objects. In: CVPR. pp. 11591–11601 (2021) 10
44. Margolin, R., Zelnik-Manor, L., Tal, A.: How to evaluate foreground maps? In: CVPR. pp. 248–255 (2014) 10
45. Mei, H., Yang, X., Wang, Y., Liu, Y., He, S.: Don't hit me! glass detection in real-world scenes. In: CVPR. pp. 3687–3696 (2020) 11, 12
46. Mei, H., Yang, X., Yu, L., Zhang, Q., Wei, X., Lau, R.W.: Large-field contextual feature learning for glass detection. IEEE Trans. Pattern Anal. Mach. Intell. (2023) 11
47. Mou, C., Wang, Q., Zhang, J.: Deep generalized unfolding networks for image restoration. In: CVPR. pp. 17399–17410 (2022) 7
48. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023) 11
49. Pang, Y., Zhao, X., Xiang, T.Z., Zhang, L., Lu, H.: Zoom in and out: A mixed-scale triplet network for camouflaged object detection. In: CVPR. pp. 2160–2170 (2022) 3
50. Silva, J., Histace, A., Romain, O., Dray, X.: Toward embedded detection of polyps in wce images for early diagnosis. Int. J. Comput. Assist. Radiol. Surg. **9**, 283–293 (2014) 10, 12
51. Skurowski, P., Abdulameer, H., Błaszczyk, J.: Animal camouflage analysis: Chameleon database. Unpublished manuscript p. 7 (2018) 10
52. Sun, K., Chen, Z., Lin, X., Sun, X., Liu, H., Ji, R.: Conditional diffusion models for camouflaged and salient object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025) 12
53. Sun, Y., Xu, C., Yang, J., Xuan, H., Luo, L.: Frequency-spatial entanglement learning for camouflaged object detection. In: ECCV. pp. 343–360 (2024) 12, 13
54. Tajbakhsh, N., Gurudu, S.R., Liang, J.: Automated polyp detection in colonoscopy videos using shape and context information. IEEE Trans. Med. Imaging **35**(2), 630–644 (2015) 10
55. Wang, W., Sun, H., Wang, X.: Lssnet: A method for colon polyp segmentation based on local feature supplementation and shallow feature supplementation. In: MICCAI. pp. 446–456. Springer (2024) 10
56. Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pvt v2: Improved baselines with pyramid vision transformer. Comput. Vis. Media **8**(3), 415–424 (2022) 11
57. Wang, X., Zhang, Z., Gao, J.: Polarization-based camouflaged object detection. Pattern Recognition Letters **174**, 106–111 (2023) 11, 13

58. Wu, H., Zhang, Z., Zhang, W., Chen, C., Liao, L., Li, C., Gao, Y., Wang, A., Zhang, E., Sun, W., et al.: Q-align: Teaching lmms for visual scoring via discrete text-defined levels. arXiv preprint arXiv:2312.17090 (2023) 8
59. Xia, B., Zhang, Y., Wang, S., Wang, Y., Wu, X., Tian, Y., Yang, W., Van Gool, L.: Diffir: Efficient diffusion model for image restoration. In: ICCV (2023) 4, 14
60. Xiao, F., Feng, J., Hu, P., Zhang, D., Xu, L., Qin, G., Li, L., He, C., Farsiu, S.: Qualiteacher: Quality-conditioned pseudo-labeling for real-world image restoration. ECCV (2026) 4
61. Xiao, F., Hu, P., Xu, L., Guo, X., Qin, G., Shen, Y., Fang, C., Zhang, R., He, C., Farsiu, S.: Beyond ground-truth: Leveraging image quality priors for real-world image restoration. CVPR (2026) 4
62. Xiao, F., Hu, S., Shen, Y., He, C.: A survey of camouflaged object detection and beyond. arXiv preprint arXiv:2408.14562 (2024) 2
63. Xu, G., Yan, J., Tang, L., Zhang, Y.: Hqg-net: Unpaired medical image enhancement with high-quality guidance. IEEE Trans. Neural Networks Learn. Syst. (2023) 2, 4, 7
64. Zhang, F., Li, Y., You, S., Fu, Y.: Learning temporal consistency for low light video enhancement from single images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4967–4976 (June 2021) 11
65. Zhang, K., Liang, J., Van Gool, L., Timofte, R.: Designing a practical degradation model for deep blind image super-resolution. In: IEEE International Conference on Computer Vision. pp. 4791–4800 (2021) 11
66. Zhu, H., Li, P., Xie, H., Yan, X., Liang, D., Chen, D., Wei, M., Qin, J.: I can find you! boundary-guided separated attention network for camouflaged object detection. In: AAAI. vol. 36, pp. 3608–3616 (2022) 12